

## RESEARCH ARTICLE

# Infrared and Visible Image Fusion Based on Autoencoder Composed of CNN-Transformer

HONGMEI WANG<sup>1</sup>, LIN LI<sup>2</sup>, CHENKAI LI<sup>1</sup>, AND XUANYU LU<sup>1</sup><sup>1</sup>School of Astronautics, Northwestern Polytechnical University, Xi'an 710072, China<sup>2</sup>Beijing Research Institute of Telemetry, Beijing 100076, China

Corresponding author: Hongmei Wang (haipw@nwpu.edu.cn)

This work was supported by the Research and Development Program of Shanxi province under Grant 2023-YBGY-232.


**ABSTRACT** Image fusion model based on autoencoder network gets more attention because it does not need to design fusion rules manually. However, most autoencoder-based fusion networks use two-stream CNNs with the same structure as the encoder, which are unable to extract global features due to the local receptive field of convolutional operations and lack the ability to extract unique features from infrared and visible images. A novel autoencoder-based image fusion network which consist of encoder module, fusion module and decoder module is constructed in this paper. For the encoder module, the CNN and Transformer are combined to capture the local and global feature of the source images simultaneously. In addition, novel contrast and gradient enhancement feature extraction blocks are designed respectively for infrared and visible images to maintain the information specific to each source images. The feature images obtained from encoder module are concatenated by the fusion module and input to the decoder module to obtain the fused image. Experimental results on three datasets show that the proposed network can better preserve both the clear target and detailed information of infrared and visible images respectively, and outperforms some state-of-the-art methods in both subjective and objective evaluation. At the same time, the fused image obtained by our proposed network can acquire the highest mean average precision in the target detection which proves that image fusion is beneficial for downstream tasks.

**INDEX TERMS** Image fusion, convolutional neural network, transformer, infrared image, visible image.

## I. INTRODUCTION

The image fusion technique is the merging of images from different scenes into a single fused image that has multiple source image features [1], [2], [3], [4], [5], [6], [7]. As mentioned in [8], “an alternative to maximize the segmentation accuracy is to jointly leverage the multimodal data to further enhance feature representations”, multimodal image fusion is one of the effective means to improve the performance of tasks such as image segmentation and target detection, et al. Various modalities of images can be combined to complete image fusion, such as infrared and visible images, infrared and SAR images, visible and SAR images, and CT and MRI images in medicine, etc. Among them, infrared images are beneficial to promote target detection and recognition ability, which can avoid the influence of environments, such

as smoke, light, rain, etc. [9]. However, it also has some shortcomings like low pixel resolution, poor contrast, insufficient texture in the background, etc. Visible images have high resolution, and can reflect rich scene information, such as texture and detail information. However, it is susceptible to environmental factors including weather, smoke, occlusion [9], and it cannot highlight targets in case of interference. Therefore the fusion technique becomes a necessary choice, which can combine the complementary advantages of both to obtain a fused image with bright targets and detailed background. Currently, infrared and visible image fusion techniques are widely used in image enhancement [10], agricultural automation [11], remote sensing detection [12], and especially in object recognition, detection and tracking [13], [14], [15]. Seal et al. [16], [17] proposed thermal and visible image fusion methods for face recognition. Experimental results demonstrated significant performance improvements in recognition over individual modality.

The associate editor coordinating the review of this manuscript and approving it for publication was Sudhakar Radhakrishnan .

In the past decades, many different image fusion methods have been proposed, which can be classified into different categories, including multiscale transform [18], [19], sparse representation [20], [21], neural network [22], [23], subspace [24], saliency [25], hybrid model [26], and other fusion methods [2], [27], [28]. In [23], the authors illustrated the basic implementation of each groups above, and pointed out that each of the groups has its strengths and shortcomings.

Although traditional fusion methods have made some progress, there still exist some problems. In general, these methods rely on manually designed feature extraction and fusion rules, which make the fusion process increasingly complex [9], [29]. Thus the lack of diversity in the features extracted in this way often leads fused images to be in low contrast, blurred textures, and artifacts in the target.

The deep learning-based image fusion models are adaptively trained to update the model parameters with the help of the learning capability of the network to form an end-to-end fusion models. Compared with traditional methods, deep learning fusion methods avoids activity level measurement and fusion rule design, which greatly reduces the influence of human factors on fusion results [29]. In addition, the deep learning methods exploit the ability of network feature extraction to fully preserve the complementary information of the source images in the fused image, which improves the quality of the fused image.

At present, image fusion methods based on deep neural network can be divided into CNN (convolutional neural network)-based methods, autoencoder-based methods, GAN (generative adversarial network)-based methods and Transformer-based methods. The image fusion networks based on autoencoder do not need manual design of fusion rule, and it has become a widely studied method nowadays. However, most of the existing encoders of autoencoder use the convolution operation which cannot fully extract the global features because of the local receptive field property. In addition, the current feature extraction subnetworks of the fusion models do not make a distinction for different source images, and complementary information is not reflected enough in fused images therefore. To solve the above problems, a novel autoencoder-based infrared and visible image fusion network combining CNN and Transformer is proposed in this paper. The contributions of this paper can be summarized as follows:

- A novel encoder composing of CNN and Transformer is established to extract the local and global information of the infrared and visible images simultaneously.
- Contrast enhancement block and gradient residual block are designed separately for infrared and visible images to maintain the complementary information of the source images.
- Extensive experiments on three datasets (TNO, OTCBVS and RoadScene) show that the proposed network can obtain the fused image containing both clear targets and rich textures, and surpasses some state-of-the-art methods, including U2Fusion, DDcGAN,

SDDGAN, DenseFuse, RFN-Nest, STDFusion and SwinFusion.

The rest of this article is organized as follows. In section II, related work is introduced. The proposed method is presented in detail in section III. Section IV conducts qualitative and quantitative comparisons of the proposed method and other state-of-the-art methods. The article is concluded in section V.

## II. RELATED WORK

In recent years, more and deep neural network models have been applied in the field of image fusion, which can be classified into four categories: CNN-based methods, GAN-based methods, autoencoder-based methods and Transformer-based methods.

### A. CNN-BASED FUSION METHODS

In 2017, Liu et al. [30] introduced CNN to the field of image fusion, where they used blurred background and foreground images to train the network and obtained a binarized weight maps. In the testing phase, the source images were combined with the weight maps to obtain a fused multi-focus image. Meanwhile, many researchers have tried to introduce the CNN modules in the traditional method, and these approaches inject rich semantic information into the fused images. For example, Li et al. [31] used the VGG19 network to further process the detailed part obtained by multiscale decomposition, thus preserving rich texture information in the fused image. Liu et al. [32] found that the features extracted by CNN could reflect the proportion of the source images to a certain extent in the fusion process, so they used the downsampling sequence of the convolutional weight maps as the fusion ratio map of the two branch downsampling sequences, avoiding artificially designed fusion strategies. Similarly, zero-phase component analysis and  $L_1$ -norm were used to obtain the weight maps reflecting the proportion of the source images [33], overcoming the problem of information loss in the process of downsampling the source images. These methods retain rich detailed information in the fused image with the help of the powerful feature extraction ability of the CNN network. However, in the above methods, CNN are mainly used in the feature extraction stage, and the traditional multiscale decomposition or fusion strategy are still applied in the fusion process. The main shortcomings of such methods include: 1) design of activity level measurement or fusion rule are still required in most CNN-based methods. 2) CNN is only used to obtain the weight maps needed for fusion and is not sufficiently involved in the whole image fusion process as a result. Also, simple weighted fusion strategy is applied as the fusion rules which will lost the information of the source images. The reason for the above problems is that the CNN-based fusion methods do not get rid of the limitations of traditional methods, and thus cannot maximize the advantages of the network itself.

## B. GAN-BASED FUSION METHODS

In 2019, Ma et al. [9] proposed a novel fusion model based on the generative adversarial network (FusionGAN), which is mainly divided into two parts: generator and discriminator. The input of the generator is the cascaded infrared and visible images, and the input of the discriminator is the image generated by the generator and the visible image. The loss function is divided into two parts: the generator loss and the discriminator loss. The generator loss is divided into adversarial loss and background loss, and the discriminator loss forces the fused image and the visible image to be more similar. The fusion network takes advantage of the GAN to preserve textures and target information for the fused image. Further, two discriminators were designed in [34] to retain the information of infrared and visible images simultaneously. Besides, researchers also designed targeted loss functions to better reserve the details of the fused image or optimize the models [35], [36], [37]. Li et al. [38] penalized the attention maps of shallow discriminators of the source and fused images, aiming to preserve more attention region of source images. The Wasserstein distance with gradient penalty were used as the loss function of the discriminator in [37] and the experimental results demonstrated that the model was more easily trained. In terms of network structure, a multi-scale attention mechanism forcing the generator to focus on the most discriminative regions of the source image was proposed in [39]. In order to make the fusion results more balanced, Ma et al. [40] designed the discriminator as a classification network, and the training process ends when the discriminator cannot distinguish between fused image and real image. Considering that the most important purpose of the fusion of infrared and visible images is to highlight the target in infrared images and preserve the background in the visible image, the literature [41], [42] used semantic segmentation methods [43] to obtain the mask image of the source images, and combined with this mask image to further obtain the 'label' for discriminator. Hong et al. [44] proposed a decoupled-and-coupled network for the hyperspectral image super-resolution task. A decoupled subnetwork was first designed by means of GANs, and then a model-driven coupled subnetwork was developed.

It provides great inspiration for the design of fusion networks for the infrared and visible images.

The GAN-based fusion method relies on the adversary between the generator and the discriminator to generate the fused images. During the training process, the 'adversarial game' can force features of the fused image to be consistent with the source images, so the fusion network can realize unsupervised training. However, this kind of fusion network is relatively difficult to be trained due to the unstable optimization of the model.

## C. AUTOENCODER-BASED FUSION METHODS

In 2018, Li and Wu [45] proposed a fusion network consisting of encoding module, fusion module, and decoding module. In the coding module, the method cascaded the

extracted feature images from each layer and input them to the next layer, which increased information flow capability. Also, the network was easier to be trained, and the final fused image retained a large amount information of source images. However, due to the lack of datasets, grayscale images from MS-COCO dataset [46] were used to train the autoencoder network. Zhao et al. [47] used infrared and visible images as training data, and combined with the hopping connection operation to highlight the information specific to source images. Liu et al. [48] proposed a fusion network with adaptive weight assignment strategy. That is to say, each transmitted feature image was assigned different weight. The fusion network was optimized and the fusion results could retain more detailed information as a result. To prevent the loss of information in the intermediate layers during the transfer process, in [49], a nested network was applied to fully extract the feature images corresponding to each convolutional layer. So the features at different scales were better retained in the fused image. Compared with CNN-based fusion models, autoencoder-based fusion models needs to design the fusion strategy, while the current fusion strategy of most networks is still simple summation, averaging, or  $L_1$ -norm [45]. Li et al. [50] designed a fusion model by training the coder, decoder and the residual fusion module separately, thus avoiding the design the fusion strategy manually. To make the network more sensitive to the target as well as the background, Ma et al. [51] introduced semantic segmentation method to obtain a binary mask that can separate the target and the background. This binary mask was then applied to the source images to obtain training data. In [52], inspired by coupled spectral unmixing, a two-stream convolutional autoencoder framework is taken as backbone to jointly decompose MS and HS data into a spectrally meaningful basis and corresponding coefficients, and a network for unsupervised hyperspectral super-resolution is proposed.

The underlying structure of the autoencoder model belongs to the CNN and the generator of the GAN can also be designed as an autoencoder structure, so the autoencoder-based fusion models are more migratory and can be more easily modified and monitored. As a result, we designed our fusion model based on autoencoder structure in this paper.

## D. TRANSFORMER-BASED FUSION METHODS

In 2017, Vaswani et al. [53] proposed a network model with a completely different principle from CNN, namely Transformer. The model was first applied to the field of natural language processing, including machine translation, sentiment classification, and word prediction. The key feature of Transformer is the self-attention mechanism, which helps the model learn the global context and enables the model to acquire remote dependencies. Driven by the great success of natural language processing, Transformer models have also been applied in computer vision tasks (Vision Transformer, ViT) [54] and achieved considerable success in many fields, such as image classification, video classification, target detection, semantic segmentation, etc.



FIGURE 1. Samples for network training.

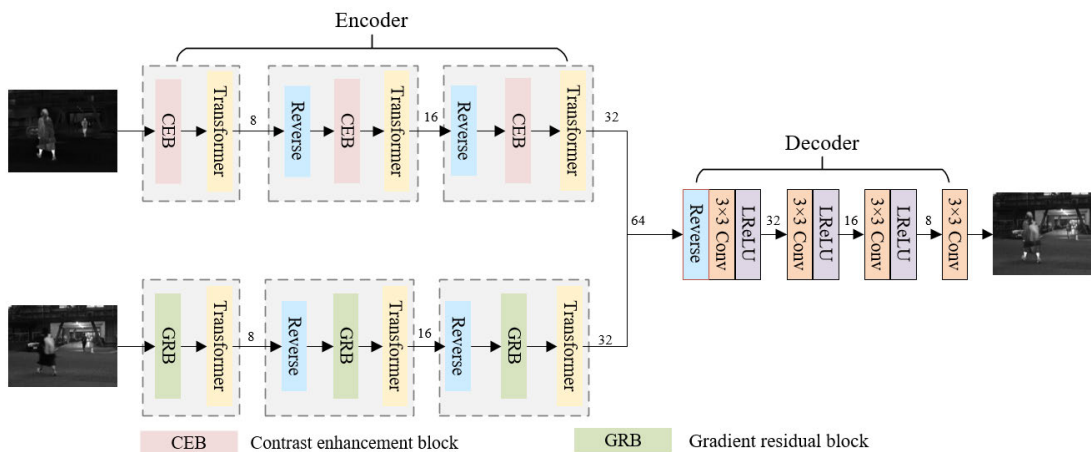


FIGURE 2. Proposed infrared and visible image fusion network.

In the field of image fusion, VS et al. [55] proposed an embedded autoencoder structure. In the fusion layer of the network, they designed spatial and Transformer fusion modules aiming to fuse local and global information. However, the fusion results were not satisfactory. In addition, Ma et al. [56] proposed a multi-task fusion network for cross-channel information interaction. They first extracted the shallow and deep feature of the source images using CNN, and later passed the feature images into the Transformer model with cross-channel information interaction to complete the interaction of source images while extracting the global information, and finally completed multi-task image fusion.

Thanks to the Transformer, a new way to extract global information from images has been gained. As a result, CNN and Transformer are combined to form an encoder in this paper, which is used to extract the local and global features of source images simultaneously, solving the shortcomings of traditional CNN-based autoencoder fusion network. In addition, based on the principle of image fusion, feature enhancement blocks have been designed separately for infrared and visible images to maintain the complementary information of the source images. The next section provides a detailed introduction to the method proposed in this article.

### III. THE PROPOSED METHOD

#### A. TRAINING DATA

A publicly available and well-aligned MSRS dataset [57] is selected as the training dataset, including 1083 pairs of

images of different scenes, whose scene targets contain people, cars, etc. The images in the dataset have a uniform size ( $640 \times 480$ ) with a bit depth of 24. To be suitable for network training, these images are processed into grayscale images and cropped the image to  $128 \times 128$  size. Some samples from MSRS dataset are shown in Fig.1, where the first four columns are the image pairs in poor-light scenes and the last four columns are the image pairs under good-light scenes.

#### B. OVERALL FRAMEWORK

The proposed fusion model consist of feature extraction, feature fusion, and feature reconstruction. The feature extraction module is divided into three stages, the CNN-based feature enhancement block and Transformer are concatenated in each stage for two streams (infrared stream and visible stream). The feature fusion module stacks the extracted features and feeds them into the decoder which is composed of three convolution layers to realize feature reconstruction and generate a fused image that contains both infrared and visible image features. The overall framework of our proposed fusion model is presented in Fig.2. Next, we will provide a detailed introduction to the CNN feature extractor designed in the encoder for infrared and visible images, as well as the Transformer module.

#### 1) CONTRAST ENHANCEMENT BLOCK (CEB) FOR INFRARED IMAGE

Contrast enhancement block is designed to enhance the contrast of the infrared image, as shown in Fig.3. The input



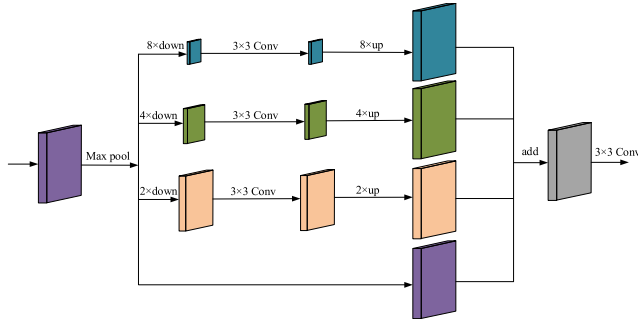


FIGURE 3. Contrast enhancement block.

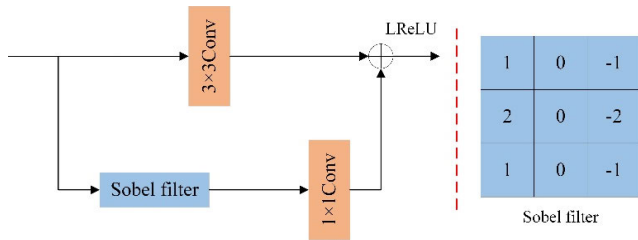


FIGURE 4. Gradient residual block.

feature image is maximally pooled (the maximum pooling operation can retain the larger pixel information in the infrared image and filter out the unimportant information) using pooling layers with step sizes of 2, 4 and 8, followed by a  $3 \times 3$  convolution operation, respectively. After which these feature images are linearly interpolated to the input feature image size and summed with the input feature image. Then, a  $3 \times 3$  convolution kernel is used to prevent artifacts in the superimposed pixels. The contrast-enhanced feature image is obtained eventually.

2) GRADIENT RESIDUAL BLOCK (GRB) FOR VISIBLE IMAGE  
Gradient residual block is designed to enhance the details of the visible image. The block adopts a residual connection mode, and its structure is shown in Fig.4. The Sobel filter is responsible for extracting the gradient information of the visible image or features. A convolutional kernel of size  $3 \times 3$  is applied as the main channel feature extractor and the activation function is Leaky ReLU. A convolutional kernel of size  $1 \times 1$  is used as the secondary channel feature extractor and the activation function is still Leaky ReLU.

### 3) TRANSFORMER MODULE

The Transformer module designed in our model is similar to the ViT model, but differs in its input and attention calculation methods. First, the two-dimensional source images are stretched to one dimensional matrix and input to the traditional Transformer model along with information such as the number of batches and the number of images per batch. In addition, inspired by SwinFusion, the feature image is divided into multiple small windows and the global attention is computed for these windows (window-based multi-head self-attention mechanism, W-MSA) to overcome the

computational complexity of traditional models. The sliding window size is set to 8, which has a larger perceptual field compared to CNN. Fig.5 illustrates the calculation process of our Transformer module. Followings are the computation process of multi-head self-attention.

For a feature image  $X \in \mathbb{R}^{M^2 \times C}$ , three learnable weight matrices  $W^Q \in \mathbb{R}^{C \times C}$ ,  $W^K \in \mathbb{R}^{C \times C}$  and  $W^V \in \mathbb{R}^{C \times C}$  are employed to project it into query  $Q$ , key  $K$ , and value  $V$ ,

$$\{Q, K, V\} = \{XW^Q, XW^K, XW^V\} \quad (1)$$

Then the attention mechanism is defined as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

where  $d_k$  is a constant value, which is convenient for finding the gradient after the Softmax operation. After the patch embedding process, the following calculations are performed,

$$\hat{f}^l = W - MSA(LN(f^{l-1})) + f^{l-1} \quad (3)$$

$$f^l = MLP(LN(\hat{f}^l)) + \hat{f}^l \quad (4)$$

where  $\hat{f}^l$  and  $f^l$  denote the output feature images of the W-MSA and MLP (Multilayer Perceptron) block, respectively.

### C. LOSS FUNCTION

In this paper, the loss function of literature [51] is introduced to make the local grayscale of the fused image similar to the source image with a larger grayscale value in the corresponding region, and to make the local gradient of the fused image similar to the source image with a larger gradient. The loss function is defined as the following equation.

$$L = \alpha L_{int} + \beta L_{texture} \quad (5)$$

where  $\alpha$  and  $\beta$  are the hyper-parameters that control the trade-off of pixel loss function  $L_{int}$  and gradient loss function  $L_{texture}$ .  $L_{int}$  and  $L_{texture}$  are defined as following respectively.

$$L_{int} = \frac{1}{HW} \|I_f - \max(I_{ir}, I_{vis})\|_1 \quad (6)$$

$$L_{texture} = \frac{1}{HW} \|\ |\nabla I_f| - \max(|\nabla I_{ir}|, |\nabla I_{vis}|) \|_1 \quad (7)$$

where  $H$  and  $W$  denote the width and height of feature images, respectively,  $I_{ir}$ ,  $I_{vis}$  and  $I_f$  denote infrared image, visible image, and fused image, respectively,  $I_{ir}$ ,  $I_{vis}$  and  $I_f$  denote gradient of infrared image, visible image, and fused image, respectively.

## IV. EXPERIMENTAL RESULTS

In this section, ablation experiments are designed first to validate the reasonableness of the model. Furthermore, to verify the advantages of the proposed model over other state-of-the-art methods, two traditional methods and seven deep learning methods are selected for subjective and objective comparisons. The objective evaluation metrics include mutual information (MI), standard deviation (SD), average gradient (AG), peak signal-to-noise ratio (PSNR), visual information

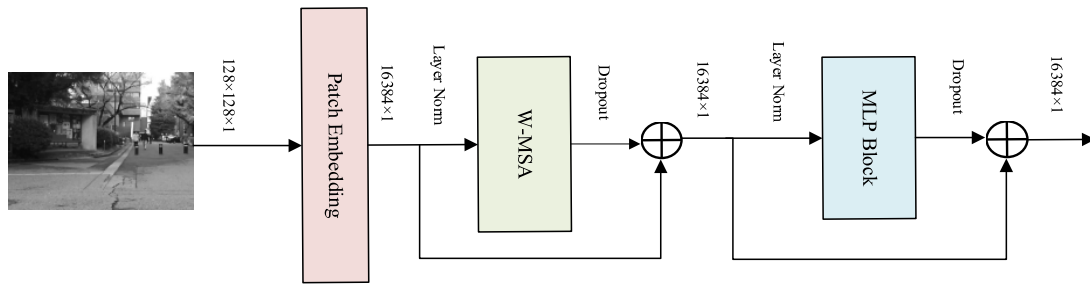


FIGURE 5. Transformer block based on W-MSA.

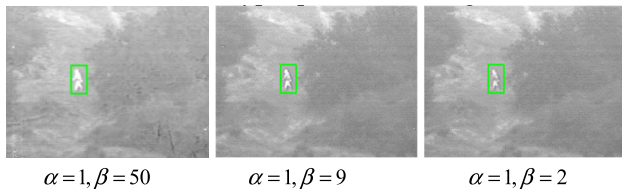


FIGURE 6. Experiments to determine the optimal hyper-parameter value.

fidelity of fusion (VIFF), and gradient-based fusion performance (QAB/F). We hope that fused images not only have high-quality visual effects, but also benefit downstream tasks, such as target detection and target recognition. Therefore, we also verify the effectiveness of target detection using the fused images in this experimental section.

#### A. MODEL CONFIGURATION

In our proposed fusion model, the learning rate is set to  $2 \times 10^{-5}$ , the decay rate is 0.99, the weight update rule is Adam, and the batch image size is set to 8, respectively. The coefficients  $\alpha$  and  $\beta$  in loss function of Eq.(5) are more important in hyper-parameters. Experiments are conducted to determine the value of these two parameters. Fig.6 presents the fusion results with different hyper-parameter setting.

From the experiments, we can find that the background texture is not rich when  $\alpha = 1$  and  $\beta = 50$ . The target is not clear when  $\alpha = 1$  and  $\beta = 2$ . The target is clear and the background is rich when  $\alpha = 1$  and  $\beta = 9$ . As a result, we conducted all experiments using  $\alpha = 1$  and  $\beta = 9$ .

The experimental operating system is Windows 10. The hardware platform is AMD Ryzen Threadripper PRO 3945WX with 4.0GHz, GPU RTX 3080, and 10G video memory. The software platform is Python 3.7, and the model is built using Pytorch to complete the training. Samples are randomly selected from three public datasets, TNO, OTCBVS, and RoadScene, as the test set.

#### B. ABLATION EXPERIMENTS

In our proposed model, different CNN feature enhancement blocks are designed for the infrared and visible channels aiming to extract the unique features of source images separately. In addition, a Transformer module is added after the CNN feature enhancement block at each stage to better extract the global features of the source images. Therefore, the purpose

of the ablation experiment is to verify the effectiveness of the CNN feature enhancement blocks and Transformer module.

Fig.7 (a) presents the fusion result using the Transformer module and contrast enhancement blocks for both channels. Fig.7 (b) presents the fusion result using the Transformer module and gradient residual blocks for both channels. Fig.7 (c) is the fusion result without using the Transformer module and Fig.7 (d) is the fusion result of our proposed model.

It can be found the visual contrast between the target and the background of the first image is better, but the details are not very rich in Fig.7 (a). Compared to the first image, the second image Fig.7 (b) has a clearer texture (shown as in the zoomed red box), but there is a ‘fault’ problem in the background. From Fig.7 (c), we can find that the overall gray value of the image is relatively high, with a lot of background missing, and the global information of the fused image is not complete enough.

Overall, the fusion result of the proposed network combines the advantages of Fig.7 (a) and Fig.7 (b), balancing the visual contrast and texture of the images with clear target and good visual effects. The objective indexes of the fused images in Fig.7 (recorded as (a), (b), (c) and (d) respectively) are shown in TABLE 1 (bold numbers represent the best performance).

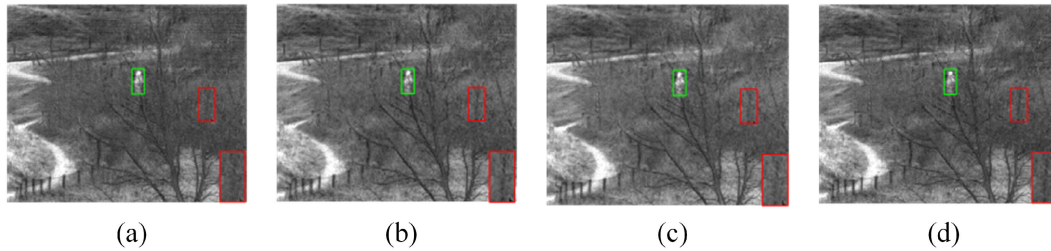
We can find that the proposed model has the highest values of MI and  $Q^{AB/F}$  indexes, so the method in this paper has the best performance in terms of information richness and local edge preservation of the fused image.

In addition, 20 images are randomly selected and the average values of the objective indexes of these 20 images are computed and show in Fig.8.

It can be found that the results of the proposed fusion model perform best, as reflected by the fact that the MI, SD, AG, and  $Q^{AB/F}$  are greater than those of other methods, the PSNR and VIFF are ranked second. In summary, the feature enhancement strategy proposed in this paper that combines contrast enhancement block, gradient residual block, and Transformer module can improve the quality of fused images.

#### C. ALGORITHM COMPARISON

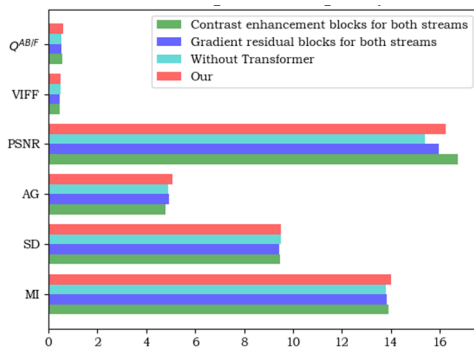
To further verify the superiority of our proposed model, nine methods are selected for subjective and objective comparison, including two traditional methods: MST\_SR [18],



**FIGURE 7. Subjective comparison of fusion results. (a) fusion result with Transformer and contrast enhancement blocks for two streams; (b) fusion result with Transformer and gradient residual blocks for two streams; (c) fusion result without Transformer module; (d) fusion result with the proposed method.**

**TABLE 1. Performance comparison between different methods in terms of six objective indexes.**

	MI	SD	AG	PSNR	VIFF	$Q^{AB/F}$
(a)	13.5606	8.0926	5.0577	<b>14.6703</b>	0.2940	0.4993
(b)	13.7125	8.1092	<b>5.2494</b>	13.9398	0.3258	0.4723
(c)	13.7024	<b>8.2029</b>	5.1412	12.9250	<b>0.3602</b>	0.4707
(d)	<b>13.7723</b>	8.1772	5.1388	13.5543	0.3199	<b>0.5157</b>



**FIGURE 8. Mean values of objective indexes for different methods.**

GTF [2]; one CNN-based method: U2Fusion [53]; two GAN-based methods: DDcGAN [30], SDDGAN [40]; three Autoencoder-based methods: DenseFuse [42], RFN-Nest [47], and STDFusion [48]; one Transformer-based method: SwinFusion [52].

1) QUALITATIVE COMPARISON

Fig.9 shows the fusion results of a pair of infrared and visible images from the TNO dataset by the above nine methods, respectively. From the infrared target perspective, the target brightness of Fig.9 (c), (d), (e), (f) and (g) is low. In addition, the target in these four images is blurry, as shown in the green box in the lower left corner of the fused image. The edges of target of Fig.9 (h) appear vague, and the edges of target of Fig.9 (i), (j) and (k) are not complete enough. Comparing with the above targets of the fused image, the targets of the fused image obtained by our proposed model are bright and clear. In addition to comparing the degree of

preservation of infrared targets, we also compare the richness of the background texture of different method, as shown in the zoomed red box in the lower right corner of the fused image. We can find that the background hierarchy is not prominent enough in Fig.9 (c), (e), (f), (g) and (i). The background textures are abundant in Fig.9 (h), (j), (k) and our proposed method. Overall, the method in this paper outperforms other methods in terms of target clarity and background texture richness.

Fig.10 show the fusion results of a pair of infrared and visible images from the OTCBVS dataset by the different methods, respectively.

It is obviously that the target brightness is low in Fig.10 (c), (d), (e), (f), and (g). The infrared target information almost can't be reflected in Fig.10 (d) and (g). Even though the targets in Fig.10 (h), (i) and (j) are bright, artifacts still exist around the target in Fig.10 (h); the target in Fig.10 (i) is incomplete; the target in Fig.10 (j) only retains the infrared image information and the texture is lost. Fig.10 (k) and our proposed method have bright targets relatively. From the region framed by the red box, we can find the texture of the tree branches in Fig.10 (e), (g), and (i) are blurred. The grayscale values of images (d), (f), (g), (h), and (i) are low and the visual effect is not satisfactory. The fused image of our method can preserve the target feature and the details of background to the greatest extent. Therefore, the fusion effect of the method in this paper is better than other comparison methods.

Next, we conducted a comparative experiment on Road-Scene dataset. The first row of Fig.11 shows a pair of infrared and visible images of a road scene from the RoadScene dataset. The green, red and blue boxes mark the target, street-lights and tree backgrounds of the fused images, respectively.

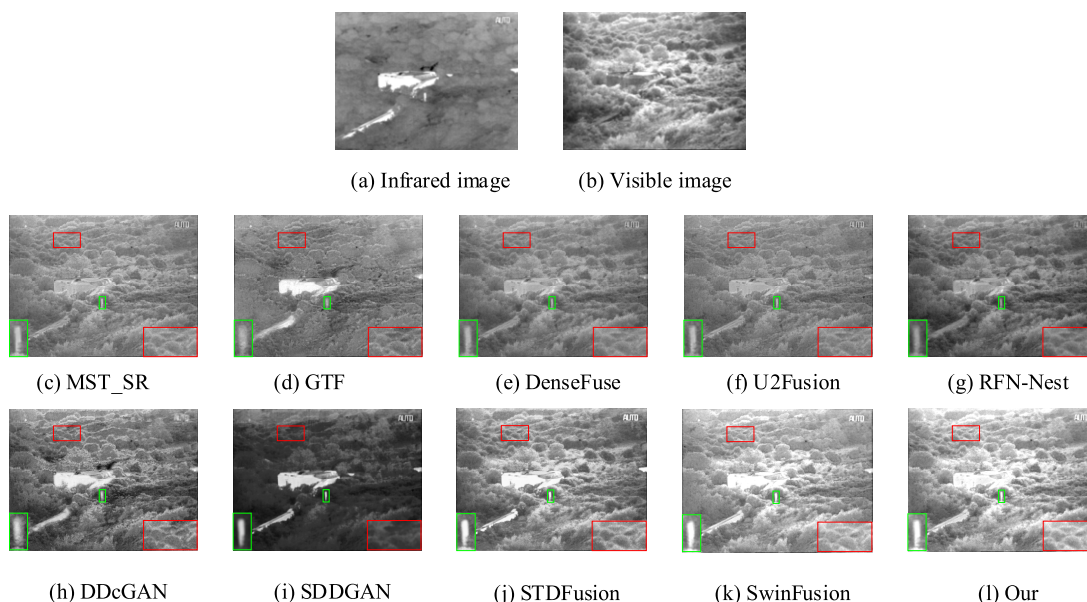


FIGURE 9. The fusion results of the sample from TNO dataset.

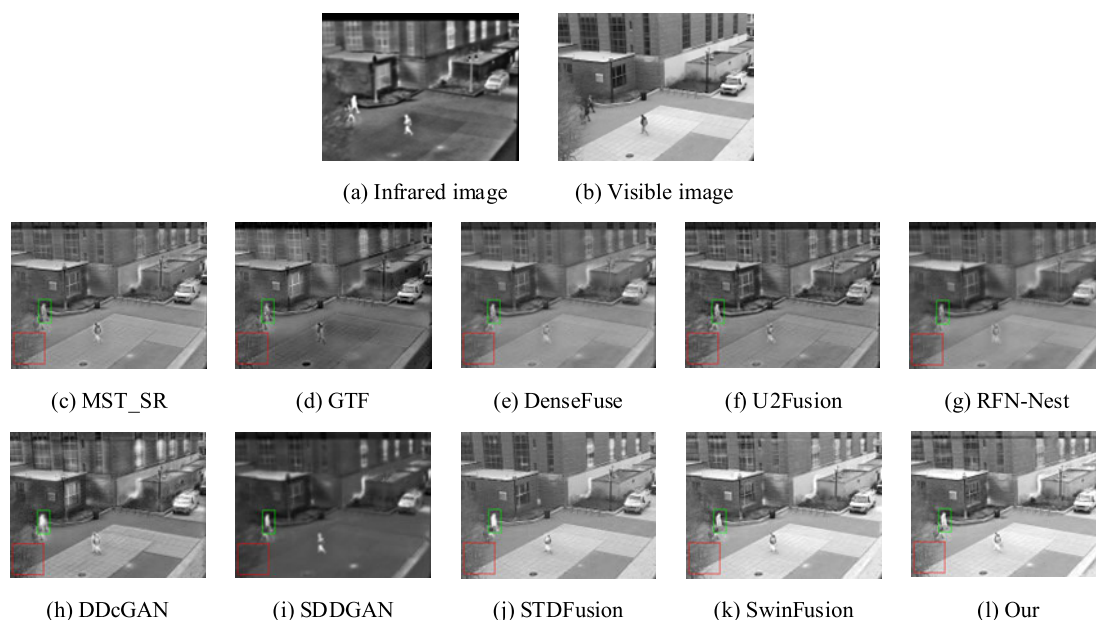


FIGURE 10. The fusion results of the sample from OTCBVS dataset.

We can find that Fig.11 (c), (e), (f), (g), and (h) have low grayscale values for the target, and the targets in Fig.11 (g), (h), and (i) are blurred. In Fig.11 (d), (h), (j) and (k), the streetlights information is incomplete. The tree backgrounds information are missing in Fig.11 (j). The tree backgrounds are blurred in Fig.11 (d), (g), (i), and (k). In Fig.11 (h), the background structures are not consistent with that of the source image. In summary, we can see that the fused image of the proposed method has bright target and plentiful background, namely, better quality.

## 2) QUANTITATIVE COMPARISON

To further validate the advantages of the proposed method in this paper, the objective evaluation metrics of each method on the above fused images are compared.

TABLE 2 records the objective evaluation metrics of fused image obtained by different methods on Fig.9. We can find that proposed method obtained three highest metrics of SD, VIFF, and  $Q^{AB/F}$ , which indicates that its fused images have bright targets, rich local information, and good visual effects. In addition, other indexes of the proposed method are at the



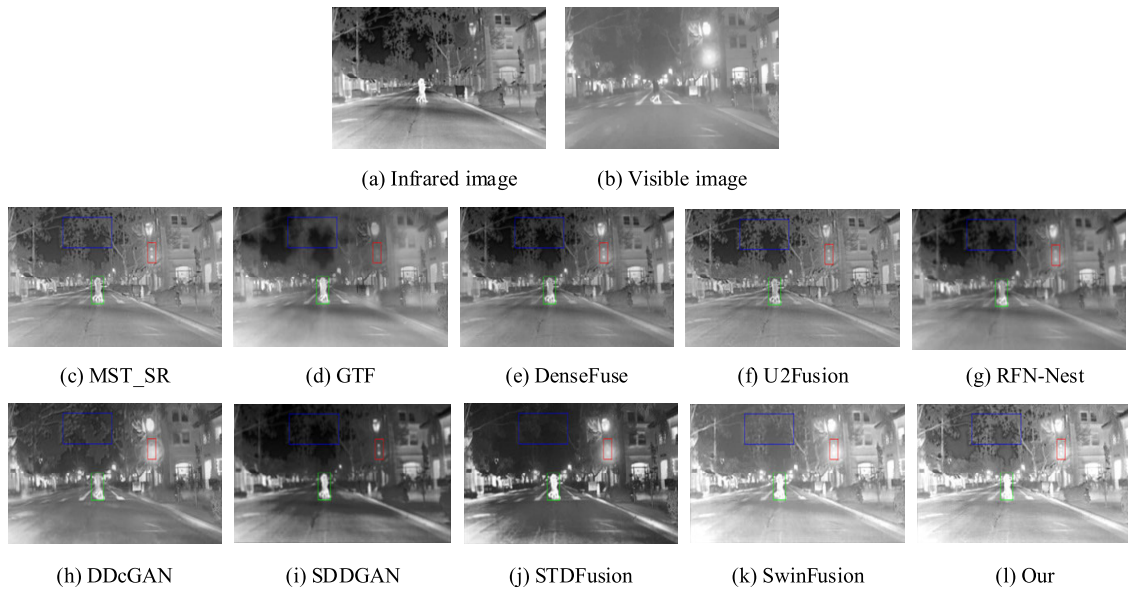


FIGURE 11. The fusion results of the sample from RoadScene dataset.

TABLE 2. Performance comparison between different methods on FIGURE 9.

	MI	SD	AG	PSNR	VIFF	$Q^{AB/F}$
MST_SR	13.6056	9.2296	4.8200	17.2963	0.2400	0.5680
GTF	13.5599	8.4677	5.2656	20.6127	0.1546	0.4872
DenseFuse	13.6332	9.3504	3.5434	17.7113	0.2639	0.3489
U2Fusion	13.1709	8.9437	4.2750	17.5764	0.2354	0.3735
RFN-Nest	14.0501	9.6343	3.1904	16.7863	0.2842	0.3313
DDcGAN	14.6560	9.4515	<b>6.8695</b>	15.0726	0.3333	0.4280
SDDGAN	13.3299	7.9004	2.0650	10.9414	0.1371	0.1196
STDFusion	<b>14.6642</b>	9.6321	6.1548	12.7968	0.3031	0.6330
SwinFusion	14.3104	9.6294	5.4706	<b>21.3936</b>	0.3288	0.4971
Our	14.6597	<b>10.0651</b>	5.6859	19.1290	<b>0.3841</b>	<b>0.6155</b>

TABLE 3. Performance comparison between different methods on FIGURE 10.

	MI	SD	AG	PSNR	VIFF	$Q^{AB/F}$
MST_SR	15.5147	10.6960	7.7727	21.2855	0.2530	0.5174
GTF	13.9555	9.7600	7.5123	21.0870	0.1632	0.4268
DenseFuse	14.2473	10.7086	5.4818	21.3639	0.2862	0.3585
U2Fusion	14.3106	10.8111	7.1326	21.4633	<b>0.2922</b>	0.4007
RFN-Nest	14.3898	10.6649	4.3975	21.1618	0.2657	0.2525
DDcGAN	14.7703	9.7284	8.2496	19.0378	0.2302	0.4037
SDDGAN	12.9774	8.6631	3.0534	21.2614	0.1653	0.1309
STDFusion	14.7488	10.1876	7.9250	20.2855	0.1664	0.5321
SwinFusion	15.0636	10.7638	8.1200	21.7800	0.2324	0.4770
Our	<b>15.1733</b>	<b>10.9592</b>	<b>8.8327</b>	<b>21.9522</b>	0.2799	<b>0.5556</b>

TABLE 4. Performance comparison between different methods on FIGURE 11.

	MI	SD	AG	PSNR	VIFF	$Q^{AB/F}$
MST_SR	14.3404	10.6224	5.3479	<b>18.4476</b>	0.5017	<b>0.6397</b>
GTF	<b>15.1163</b>	10.5803	3.1869	18.2426	0.3151	0.3377
DenseFuse	14.6365	10.8247	4.5066	18.2426	0.5619	0.5404
U2Fusion	14.3202	10.7503	5.4099	17.9279	0.5047	0.5690
RFN-Nest	14.7527	<b>10.9460</b>	2.9120	18.3478	0.4601	0.2809
DDcGAN	14.4924	9.4595	4.5572	15.4262	0.3332	0.3244
SDDGAN	14.7761	9.6421	3.3221	15.5799	0.4372	0.2241
STDFusion	14.6468	8.7094	5.3883	10.2598	0.4942	0.3787
SwinFusion	14.4450	10.6648	4.6048	17.0968	0.4541	0.3887
Our	14.7907	10.6445	<b>6.3884</b>	15.1769	<b>0.6772</b>	0.6054

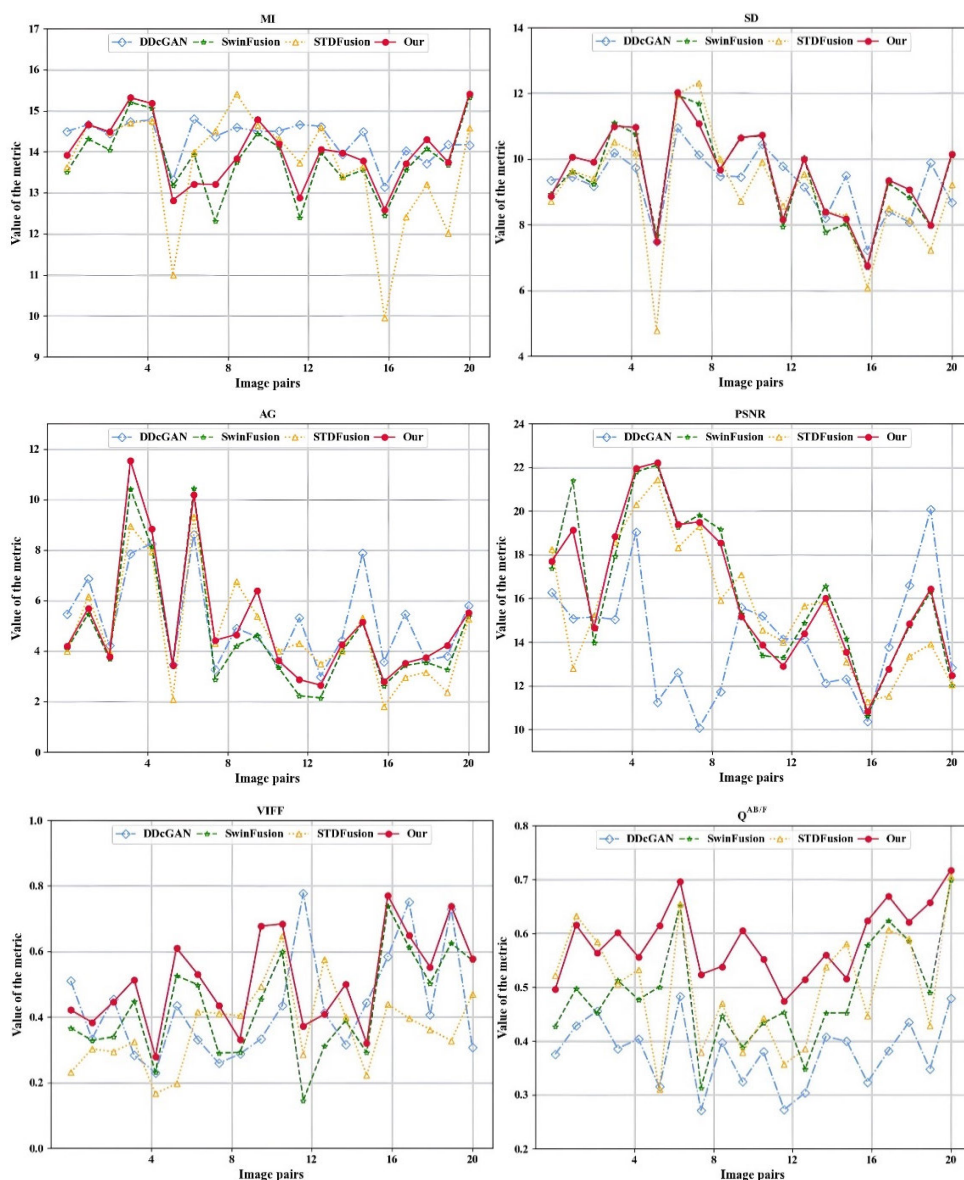


FIGURE 12. Objective performance of different fusion methods on the test set.

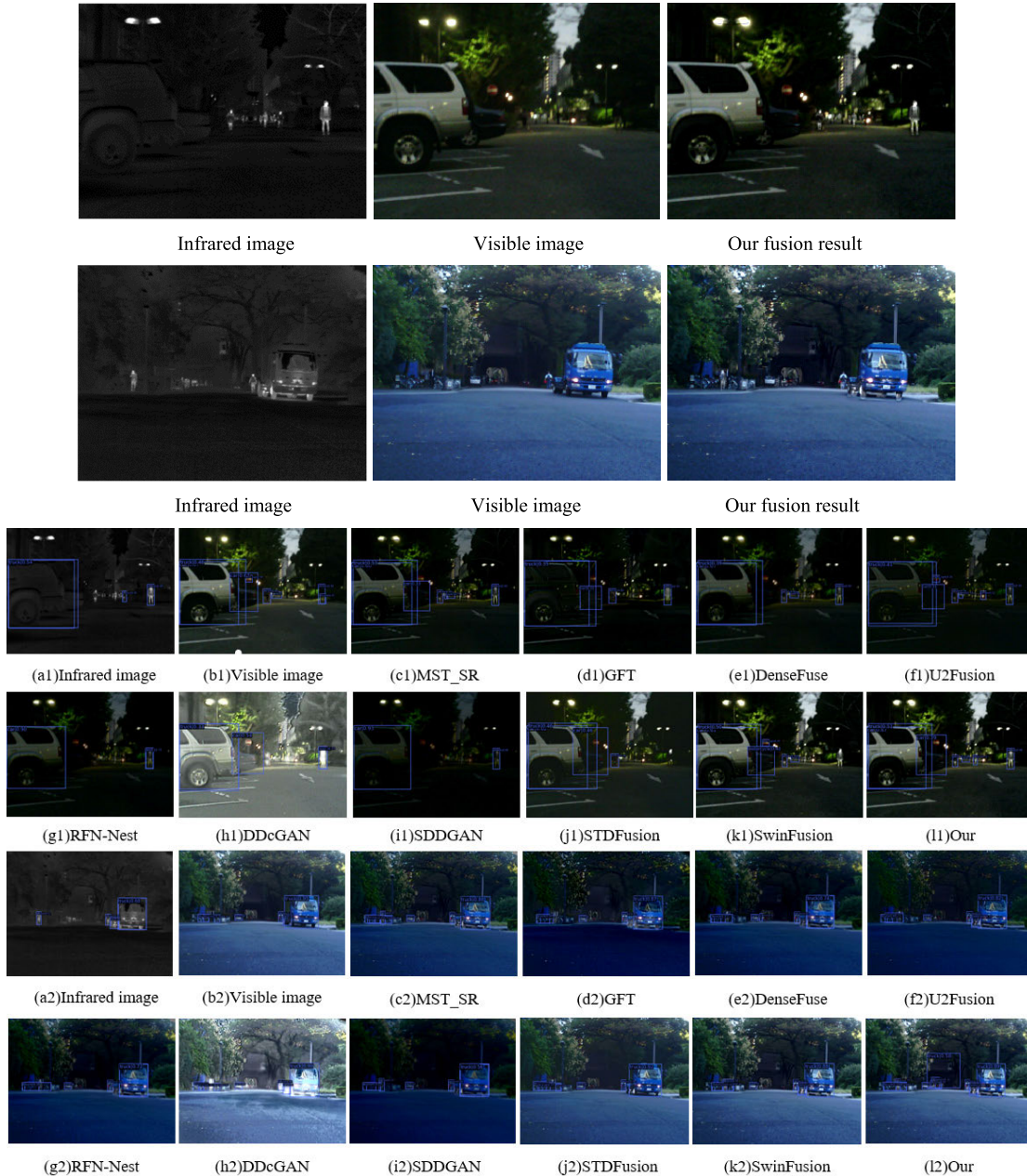


FIGURE 13. Fusion results and target detection results for source images and fused images.

top among all the methods, for example MI value ranks in the second position, which is only 0.001 smaller than the STDFusion method.

TABLE 3 records the objective evaluation metrics of the fused image obtained by different methods on Fig.10. In terms of objective metrics, the method in this paper similarly achieves the optimal metrics except for the VIFF value, which ranks second. Consistent with the subjective evaluation, the information content, texture, contrast, and visual effect of the fused images of the proposed method are better than the latest state-of-the-art methods.

TABLE 4 records the objective evaluation metrics of the fused image obtained by different methods on Fig.11. We can

find that our proposed method performs best in AG and VIFF metrics, and the values of both MI and  $Q^{AB/F}$  metrics rank second, which demonstrate the fused image of the proposed method is of high quality and objectively evaluated well.

In addition to the above single image metrics comparison, 20 pairs of infrared and visible images are randomly selected from the test set to objectively compare the performance of the fused images obtained by different methods (in order to present the results more clearly, we selected three deep neural network fusion methods with better performance and compared them with the proposed method), as shown in Fig.12. It can be noticed that the proposed method outperforms the other methods in SD, AG, VIFF, and  $Q^{AB/F}$  metrics overall,



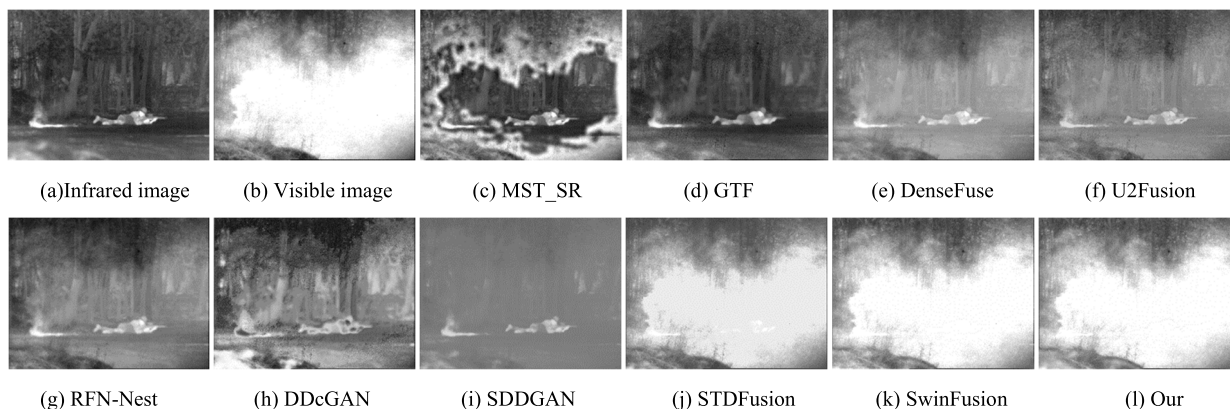


FIGURE 14. Failure case of the fusion methods.

TABLE 5. Computation Time of different fusion methods.

	Computation Time/s
MST_SR	0.3147
GTF	0.725
DenseFuse	0.7892
U2Fusion	0.7820
RFN-Nest	0.2660
DDcGAN	0.4251
SDDGAN	0.1830
STDFusion	0.5623
SwinFusion	2.5250
Our	<b>0.4219</b>

and 16 fused images rank among the top  $Q^{AB/F}$  metric values of the compared methods.

At the same time, the average computation time of different methods is provided in Table 5. We can find that our proposed fusion network runs faster, especially compared to other deep neural networks, which is beneficial for engineering applications.

In summary, the objective evaluation of the fused images of the proposed method is better than other methods overall.

#### D. COMPARATIVE EXPERIMENT OF TARGET DETECTION BETWEEN THE SOURCE IMAGES AND FUSED IMAGES

To demonstrate the value of the fused images obtained by the proposed method in the application, the target detection results of the infrared images, visible images, fused images by other fusion methods and our method are compared. Fifty image pairs for target detection are selected from MFNet dataset which contain some urban scenes, and some important targets such as people and cars are manually labeled in this paper. The infrared images, visible images, and the fused images by different methods are fed to the target detection network Faster R-CNN [58] separately to obtain the corresponding detection maps. The detection performance is measured using the mean average accuracy (mAP) and the detection results are shown in TABLE 6.

TABLE 6. Target detection accuracy of different fusion methods.

	mAP
Infrared image	0.494
Visible image	0.561
MST_SR	0.773
GTF	0.725
DenseFuse	0.699
U2Fusion	0.788
RFN-Nest	0.667
DDcGAN	0.654
SDDGAN	0.606
STDFusion	0.712
SwinFusion	0.730
Our	<b>0.800</b>

Obviously, the detection accuracy using source images is relatively low, so the necessity of image fusion has been verified. Secondly, among all the fusion methods, the detection accuracy of the proposed method is the highest, so the fusion method proposed in this paper is more favorable for application compared to other fusion methods.

In addition, Fig. 13 provides two examples to illustrate the advantages of the fusion method in this paper for the target detection task.

For the first scene, two people are correctly detected in each of the source infrared and visible images, and one truck is mistakenly detected as the car in the infrared image. For the RFN-Nest, DDcGAN, SDDGAN, STDFusion, and Swin-Fusion methods, only one person are detected in the fused image. For the MST\_SR, GTF and DenseFuse methods, only two persons are detected in the fused image. For our proposed fusion model, three people are detected and category of the targets (car or truck) is correctly identified for the fused image. In the second scene, only one person at different location is detected in the infrared image and the visible image separately, while two people can be simultaneously detected for the fusion result of the proposed method. Among the compared methods, all the methods except MST\_SR and



SDDGAN methods cannot detect the target in the infrared image. In summary, the fusion results of the proposed method have great application value in other tasks, such as target detection.

### E. FAILURE CASES OF THE PROPOSED METHOD

Although the proposed network is effective in most case, there are some failure cases with which large area of smoke appears in visible images. Fig.14 shows a failure case of the proposed method.

### V. CONCLUSION

A novel CNN-Transformer architecture based autoencoder for infrared and visible image fusion is proposed. On the one hand, to address the problem that most encoders in autoencoder networks use CNN, which is not sensitive to the global information, Transformer is introduced and combined with CNN to form the encoder which can retain both local and global information, and improve the quality of the fused image therefore. On the other hand, to address the problem that current encoders uses Siamese networks and fail to adequately extract unique features from infrared and visible images, a contrast enhancement block for infrared image and gradient residual block for visible image are designed respectively.

Compared with other state-of-the-art methods, the proposed fusion method in this paper can obtain the fused images with good subjective and objective evaluations, runs faster and is more conducive to downstream tasks such as target detection.

Although the proposed fusion network is effective in most cases, the fusion results are not ideal in situations where there is large smoke interference in visible images. In future research, we will try to solve these problems by improving the network structure and loss function.

In addition, the purpose of pixel level fusion is not only to obtain high-quality fused images, but also to facilitate other tasks including object detection and recognition. Subsequently, multi-task neural networks can be developed to efficiently complete tasks such as object detection while obtaining high-quality fused images.

### REFERENCES

- [1] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.
- [2] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016.
- [3] X. Ji and G. Zhang, "Image fusion method of SAR and infrared image based on curvelet transform with adaptive weighting," *Multimedia Tools Appl.*, vol. 76, no. 17, pp. 17633–17649, Sep. 2017.
- [4] H. Li, Y. T. Zhou, and R. Chellappa, "SAR/IR sensor image fusion and real-time implementation," in *Proc. Conf. Rec. 29th Asilomar Conf. Signals, Syst. Comput.*, Oct. 1995, pp. 1121–1125.
- [5] Y. Ye, B. Zhao, and L. Tang, "SAR and visible image fusion based on local non-negative matrix factorization," in *Proc. 9th Int. Conf. Electron. Meas. Instrum.*, Aug. 2009, pp. 4-263–4-266.
- [6] M. A. Ali and D. A. Clausi, "Automatic registration of SAR and visible band remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jun. 2002, pp. 1331–1333.
- [7] K. Parmar, R. K. Kher, and F. N. Thakkar, "Analysis of CT and MRI image fusion using wavelet transform," in *Proc. Int. Conf. Commun. Syst. Netw. Technol.*, May 2012, pp. 124–127.
- [8] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, "Multimodal GANs: Toward crossmodal hyperspectral–multispectral image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5103–5113, Jun. 2021.
- [9] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.
- [10] L. Bai, W. Zhang, X. Pan, and C. Zhao, "Underwater image enhancement based on global and local equalization of histogram and dual-image multi-scale fusion," *IEEE Access*, vol. 8, pp. 128973–128990, 2020.
- [11] V. I. Adamchuk, R. V. Rossel, and K. A. Sudduth, *Sensor Fusion for Precision Agriculture* (Sensor Fusion-Foundation and Applications). Rijeka, Croatia: InTech, 2011, pp. 27–40.
- [12] Z. Wang, G. Li, and X. Jiang, "Flood disaster area detection method based on optical and SAR remote sensing image fusion," *J. Radar*, vol. 9, no. 3, pp. 539–553, 2020.
- [13] M. Rashid, M. A. Khan, M. Alhaisoni, S.-H. Wang, S. R. Naqvi, A. Rehman, and T. Saba, "A sustainable deep learning framework for object recognition using multi-layers deep features fusion and selection," *Sustainability*, vol. 12, no. 12, p. 5037, Jun. 2020.
- [14] T. Cong, L. Yongshun, Y. Hua, Y. Xing, and L. Yuan, "Decision-level fusion detection for infrared and visible spectra based on deep learning," *Infr. Laser Eng.*, vol. 48, no. 6, 2019, Art. no. 626001.
- [15] Y. Shen, "RGB-T bimodal twin tracking network based on feature fusion," *J. Infr. Millim. Waves*, vol. 50, no. 3, 2021, Art. no. 20200459.
- [16] A. Seal, D. Bhattacharjee, M. Nasipuri, C. Gonzalo-Martin, and E. Menasalvas, "Fusion of visible and thermal images using a directed search method for face recognition," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 4, Apr. 2017, Art. no. 1756005.
- [17] A. Seal and C. Panigrahy, "Human authentication based on fusion of thermal and visible face images," *Multimedia Tools Appl.*, vol. 78, no. 21, pp. 30373–30395, Nov. 2019.
- [18] X. Yang, T. Tong, and S. Y. Lu, "Fusion of infrared and visible images based on multi-features," *Opt. Precis. Eng.*, vol. 22, no. 2, pp. 489–496, 2014.
- [19] J. Chen, X. Li, L. Luo, X. Mei, and J. Ma, "Infrared and visible image fusion based on target-enhanced multiscale transform decomposition," *Inf. Sci.*, vol. 508, pp. 64–78, Jan. 2020.
- [20] B. Yang and S. Li, "Multifocus image fusion and restoration with sparse representation," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 884–892, Apr. 2010.
- [21] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, Jul. 2015.
- [22] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 953–961.
- [23] C. Panigrahy, A. Seal, and N. K. Mahato, "Parameter adaptive unit-linking dual-channel PCNN based infrared and visible image fusion," *Neurocomputing*, vol. 514, pp. 21–38, Dec. 2022.
- [24] Z. Fu, X. Wang, J. Xu, N. Zhou, and Y. Zhao, "Infrared and visible images fusion based on RPCA and NSCT," *Infr. Phys. Technol.*, vol. 77, pp. 114–123, Jul. 2016.
- [25] A. Wang and M. Wang, "RGB-D salient object detection via minimum barrier distance transform and saliency fusion," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 663–667, May 2017.
- [26] X. R. Cui, T. Shen, and J. L. Huang, "Infrared and visible image fusion based on BEMD and improved visual saliency," *Infr. Technol.*, vol. 42, no. 11, p. 1061, 2020.
- [27] S. Rajkumar and P. C. Mouli, "Infrared and visible image fusion using entropy and neuro-fuzzy concepts," in *Proc. ICT Crit. Infrastruct., 48th Annu. Conv. Comput. Soc. India*, vol. 1. Cham, Switzerland: Springer, 2014, pp. 93–100.
- [28] J. Zhao, G. Cui, X. Gong, Y. Zang, S. Tao, and D. Wang, "Fusion of visible and infrared images using global entropy and gradient constrained regularization," *Infr. Phys. Technol.*, vol. 81, pp. 201–209, Mar. 2017.
- [29] C. Sun, C. Zhang, and N. Xiong, "Infrared and visible image fusion techniques based on deep learning: A review," *Electronics*, vol. 9, no. 12, p. 2162, Dec. 2020.

- [30] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Inf. Fusion*, vol. 36, pp. 191–207, Jul. 2017.
- [31] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2705–2710.
- [32] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 16, no. 3, May 2018, Art. no. 1850018.
- [33] H. Li, X.-J. Wu, and T. S. Durrani, "Infrared and visible image fusion with ResNet and zero-phase component analysis," *Infr. Phys. Technol.*, vol. 102, Nov. 2019, Art. no. 103039.
- [34] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.
- [35] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, and J. Jiang, "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, Feb. 2020.
- [36] J. Xu, X. Shi, S. Qin, K. Lu, H. Wang, and J. Ma, "LBP-BEGAN: A generative adversarial network architecture for infrared and visible image fusion," *Infr. Phys. Technol.*, vol. 104, Jan. 2020, Art. no. 103144.
- [37] J. Li, H. Huo, K. Liu, and C. Li, "Infrared and visible image fusion using dual discriminators generative adversarial networks with Wasserstein distance," *Inf. Sci.*, vol. 529, pp. 28–41, Aug. 2020.
- [38] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, "AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks," *IEEE Trans. Multimedia*, vol. 23, pp. 1383–1396, 2021.
- [39] J. Li, H. Huo, C. Li, R. Wang, C. Sui, and Z. Liu, "Multigrained attention network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.
- [40] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.
- [41] J. Hou, D. Zhang, W. Wu, J. Ma, and H. Zhou, "A generative adversarial network for infrared and visible image fusion based on semantic segmentation," *Entropy*, vol. 23, no. 3, p. 376, Mar. 2021.
- [42] H. Zhou, W. Wu, Y. Zhang, J. Ma, and H. Ling, "Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network," *IEEE Trans. Multimedia*, vol. 25, pp. 635–648, 2023.
- [43] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801–818.
- [44] D. Hong, J. Yao, D. Meng, N. Yokoya, and J. Chanussot, "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," 2022, *arXiv:2205.03742*.
- [45] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.
- [46] T. Y. Lin, M. Maire, and S. Belongie, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.
- [47] Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, and J. Zhang, "DIDFuse: Deep image decomposition for infrared and visible image fusion," 2020, *arXiv:2003.09210*.
- [48] L. Liu, M. Chen, M. Xu, and X. Li, "Two-stream network for infrared and visible images fusion," *Neurocomputing*, vol. 460, pp. 50–58, Oct. 2021.
- [49] H. Li, X.-J. Wu, and T. Durrani, "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9645–9656, Dec. 2020.
- [50] H. Li, X.-J. Wu, and J. Kittler, "RFN-nest: An end-to-end residual fusion network for infrared and visible images," *Inf. Fusion*, vol. 73, pp. 72–86, Sep. 2021.
- [51] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, "STDFusionNet: An infrared and visible image fusion network based on salient target detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.
- [52] J. Yao, D. F. Hong, J. Chanussot, D. Y. Meng, X. X. Zhu, and Z. B. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 208–224.
- [53] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst., 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1–15.
- [54] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [55] V. Vs, J. M. J. Valanarasu, P. Oza, and V. M. Patel, "Image fusion transformer," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 3566–3570.
- [56] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via Swin transformer," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.
- [57] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.
- [58] S. Ren, K. He, and R. Girshick, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1–9.



**HONGMEI WANG** was born in 1977. She received the B.S. degrees from Northwest Normal University, in 1999, and the M.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, Shaanxi, China, in 2002 and 2005, respectively.

From 2012 to 2013, she was a Visiting Scholar with Ryerson University, Toronto, Canada. She is currently a Professor with the School of Astronautics, Northwestern Polytechnical University. Her research interests include image fusion and pattern recognition.



**LIN LI** was born in 1995. He received the B.S. degree in mechanical design, manufacture, and automation from the North China University of Water Resources and Electric Power, China, in 2019, and the master's degree in electronic information from Northwestern Polytechnical University, China, in 2023. His research interests include image processing and deep learning.



**CHENKAI LI** born in 2000. He received the B.S. degree in aerospace engineering from Northwestern Polytechnical University, Xi'an, Shaanxi, China, in 2022, where he is currently pursuing the master's degree with the School of Astronautics. His research interests include remote sensing image processing and deep learning.



**XUANYU LU** was born in 1999. He received the B.S. degree in guidance, navigation, and control technology from Northwestern Polytechnical University, China, in 2021, where he is currently pursuing the degree with the School of Astronautics. His research interests include image processing, feature fusion, and deep learning.

...